DOCUMENT RESUME

ED 467 815 TM 034 359

AUTHOR Schnipke, Deborah L.; Scrams, David J.
TITLE Item Theft in a Continuous-Testing Environment: What Is the Extent of the Danger? Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION Law School Admission Council, Princeton, NJ.
REPORT NO LSAC-R-98-01
PUB DATE 1999-05-00
NOTE 19p.
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS *Cheating; Item Banks; Simulation; *Test Items
IDENTIFIERS *Inflated Scores; *Test Security

ABSTRACT

This study explored the dangers of item theft in terms of impact on test quality and fairness. Simulations were used to explore the impact of item theft on test taker scores. In the simulation, a simulated, organized group of "thieves" took the test, memorized the items received, and distributed the items to future test takers. Impact was explored for varying numbers of thieves and for thieves of two ability levels. Results show that test takers' scores can be largely inflated when test takers have prior knowledge of items. Who gets inflated scores depends on the test taker's ability, and the number and difficulty of the stolen items. Although the results may be cause for caution, concerns raised by this study must be considered in the proper context. This study investigated one simple design with only one item pool. Other designs may provide more opportunities to counteract the threat of item theft. Possible solutions to the problem are discussed, and further research is planned to evaluate these possibilities. (Contains 1 table, 5 figures, and 28 references.) (Author/SLD)

TM

Item Theft in a Continuous-Testing Environment: What is the Extent of the Danger?

Deborah L. Schnipke
Law School Admission Council
and
David J. Scrams
Educational Testing Service

Law School Admission Council
Computerized Testing Report 98-01
May 1999

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

Computerized test administration offers a number of advantages over traditional paper-and-pencil administrations, but some of these advantages may bring additional concerns as well. One often-mentioned advantage is the potential for continuous or on-demand testing. That is, test takers may have the freedom to schedule a test administration at a date and time convenient for the test taker rather than being required to take a group-administered test at one of a few prescheduled administrations spread throughout the year. The initial allure of continuous testing may be offset by the added risk to test security. The threat of one test taker relaying items to a friend may be a relatively minor issue, but the possibility of organized, large-scale item theft is far more serious. This issue was emphasized recently when Kaplan Educational Centers demonstrated that a small number of test takers could memorize items during a test administration and effectively memorize a significant portion of items from the computerized Graduate Record Examinations.

As the Law School Admission Council (LSAC) considers computerizing the Law School Admission Test (LSAT), the concerns that accompany the benefits of computerization must be addressed. The present study explores the dangers of item theft in terms of impact on test quality and fairness.

In this preliminary investigation, simulations were used to explore the impact of item theft on test taker scores. These simulations demonstrate how estimates of test taker ability are affected when test takers have access to some of the items in the item pool before they take the test. This access is facilitated by a simulated organized group of "thieves" who take the test, memorize the items received, and distribute these items to future test takers. Impact is explored for varying numbers of thieves as well as for thieves of two general ability levels: regular thieves who are similar in ability to average test takers, and professional thieves of substantially greater ability.

Results show that when regular thieves provided the stolen items, all but the highest ability test takers received inflated ability estimates. When professional thieves provided stolen items, some low-ability test takers were helped tremendously and received ability estimates at the top of the ability range. All test takers who had relatively high abilities received a substantial number of stolen items and also received ability estimates at the top of the ability range when professional thieves provided the stolen items.

Although the results may be cause for caution, concerns raised by the current work must be considered in the proper context. First, items and even complete test forms are occasionally stolen from current paper-and-pencil test administrations. Future test takers do not know when stolen pretest items will appear again, so the benefit of memorizing stolen items is reduced. The real problem of item theft is a concern introduced by continuous testing, not by computer administration. In a continuous testing environment, memorizing stolen items is of potentially substantial benefit if the item pool still contains the stolen items. A computer-administered test need not allow continuous testing, however, if all test takers can be accommodated with a set number of prescheduled test administrations. Second, the present study investigated only a very simple test design: a fully adaptive test with constraints only on item exposure. Additionally, only one item pool was used. Other designs may provide more opportunities to counteract the threat of item theft. One promising alternative is the multistage-testlet design currently being considered for a potential computerized LSAT administration. This and other possible solutions to the problem, such as item pool swapping, are discussed, and further research is planned to evaluate these possibilities.

Abstract

The present study explores the dangers of item theft in terms of impact on test quality and fairness. Simulations were used to explore the impact of item theft on test taker scores. Results show that test takers' scores can be largely inflated when test takers have prior knowledge of items. Who gets inflated scores depends on the test taker's ability, and the number and difficulty of the stolen items. Although the results may be cause for caution, concerns raised by the current work must be considered in the proper context. The present study investigated one simple test design with only one item pool. Other designs may provide more opportunities to counteract the threat of item theft. Possible solutions to the problem are discussed, and further research is planned to evaluate these possibilities.

Introduction

Computerized test administration provides a host of benefits when compared to traditional paper-and-pencil administrations (Wainer, et al., 1990). Some potential benefits include immediate scoring, novel item types, collection of response times along with response choices, and the potential for continuous or on-demand testing. Additional benefits are offered by adaptive test administrations. In such administrations, items are selected at least partially on the basis of ongoing estimates of test taker ability. Thus, the test "adapts" to the test taker's performance. High-ability test takers tend to receive more difficult items, and low-ability test takers tend to receive easier items. Adaptive testing offers the added benefit of equal test precision with fewer items and the accompanying benefit of shorter test administrations. These benefits are important justifications for replacing a traditional paper-and-pencil test with a computer-based test (CBT) or a computerized adaptive test (CAT).

Unfortunately, these benefits carry potential costs as well. Immediate scoring may be desirable, but it complicates the handling of flawed items (Potenza & Stocking, 1997). Novel item types may tap new and important skills, but item development may be more expensive and item scoring may be more problematic. Response times may provide useful information about test takers, but use of that information may require development of new testing models (Roskam, 1997; Thissen, 1983; Verhelst, Verstralen, & Jansen, 1997), as well as changes in item-banking (Schnipke & Scrams, 1999) and item-selection procedures (van der Linden, Scrams, & Schnipke, 1999). These costs are surmountable, but research is necessary to arrive at reasonable solutions. Working in this vein, the present work considers one of the costs associated with the benefit of continuous testing; namely, the risk of item theft.

Test security is a fundamental concern for all large-scale assessments. If a test is administered simultaneously to all test takers on each of several scheduled administrations, test security can be tightly controlled by maintaining the security of the items before the administration and limiting the use of items that have appeared on previously administered test forms. This approach to test security is more difficult in the continuous-testing context. If a single CBT test form or a single CAT item pool is used continuously, test takers may be able to memorize items and provide these items to future test takers. Such behavior may undermine test quality because some test takers may respond correctly to items only because they have seen the items outside the testing context. In such cases, the test taker could receive an inappropriately high score.

Although the threat of one test taker relaying items to a friend may be a relatively minor issue, the possibility of organized, large-scale item theft is far more serious. This issue was emphasized recently when Kaplan Educational Centers demonstrated that a small number of test takers could memorize items during administration and effectively memorize a significant portion of items from the computerized Graduate Record Examinations (Honan, 1995). The present work is an attempt to quantify the threat to test quality and fairness imposed by the possibility of organized item theft.

Although no research is currently available that directly concerns the impact of item theft, Stocking, Ward, and Potenza (1997) addressed the related issue of including disclosed items in a CAT item pool. Disclosed items are those items that have previously appeared on paper-and-pencil test forms or in CAT item pools and have been made publicly available. Although disclosure is mandated by the New York Education Law, Title 1, Article 7-A, Section 342, most testing programs, including the LSAT, would probably disclose items on a regular basis as a matter of sound public policy. Potential test takers often use disclosed items to practice for tests. Thus, test takers may have been exposed to disclosed items before an operational test administration. If disclosed items were included on operational tests, test takers' performance on such items could overestimate their true ability.

Stocking, Ward, and Potenza (1997) explored the impact of using disclosed items through a series of CAT simulations. Either 10% or 20% of items were identified as disclosed. During simulations, any disclosed item presented to a test taker was answered correctly. Thus, Stocking, Ward, and Potenza considered the worst case scenario in which all test takers had memorized the correct answers to all disclosed items. The number of disclosed items that could appear in a test taker's test was constrained in one of two ways: either the maximum number that could appear was constrained or the exact number of disclosed items appearing on each test was set to a constant. The first constraint is more lenient because fewer disclosed items might appear in a given test taker's test.

Impact was evaluated in terms of estimated test reliability. Using disclosed items had the expected effect of decreased test reliability, and using 20% of disclosed items reduced reliability more than using 10%. The increase in test scores produced by using disclosed items was also investigated. As expected, use of disclosed items tended to increase test scores. Increases were larger when more disclosed items appeared (exact number appearing was specified rather than maximum number appearing), and using 20% of disclosed items resulted in larger score increases than using 10% of disclosed items. Use of disclosed items helped low-scoring test takers most and high-scoring test takers least due to the restriction of range for high-scoring test takers. Overall, the impact of using 10% of disclosed items could be considered small in terms of Cohen's definition of practical significance (1988), but use of 20% of disclosed items resulted in larger effects that might be considered undesirable.

Stocking, Ward, and Potenza's (1997) results are encouraging for testing companies that wish to reduce the item-production demands of a CAT by including disclosed items, but their results don't speak directly to the issue of item theft. When using disclosed items, test developers can tightly control the number of disclosed items appearing on any particular test (as was done by Stocking, Ward, & Potenza), but when items are memorized by test takers, test developers are unlikely to know which items are essentially disclosed. The present work is designed to fill that gap by quantifying the impact of item theft.

Impact is evaluated through a series of simulations in which both the item theft and the "contaminated" testing are simulated. The simulations are grounded in a simple scenario. An unscrupulous individual or group of individuals elicit(s) items from item thieves and distributes these items to future test takers. These coached test takers diligently memorize all stolen items. When a coached test taker encounters a stolen item, the item is always answered correctly. Item thieves are assumed to be of two types: regular thieves and "professional" thieves. Regular thieves are test takers who agree to steal items while they take the test. Regular thieves are similar in ability to other test takers taking the test. Professional thieves tend to be of higher ability than the average test taker, and thus tend to receive (and hence steal) the most difficult items in the item pool. Both regular and professional thieves are assumed to be able to steal items with perfect accuracy. This assumption is deemed reasonable because thieves could wear hidden cameras or use other technology to steal items (Colton, 1997). The details of the simulation design are presented next.

Method

The simulation design is straightforward, and consists of four parts: (1) test design and simulation of the item bank, (2) simulation of item theft, (3) simulation of test administration to coached test takers, and (4) replication. Each part is described separately.

Test Design and Simulation of the Item Bank

A basic CAT was selected as the context for this initial examination of the impact of item theft on computerized testing. Item selection is based on maximum item information along with statistical item-exposure control (Hetter & Sympson, 1997; Sympson & Hetter, 1985). Content constraints were not implemented in this study. Final ability estimates were Bayes modal estimates (Lord, 1980; also called MAP estimates for "mode of *a posteriori* distribution").

A 500-item item bank was simulated on the basis of a 3-parameter logistic (3PL) item-response-theory (IRT) model. Item discrimination parameters (a's) were selected randomly from a normal distribution with a mean of 1.0 and a standard deviation of 0.2. Item difficulty parameters (b's) were selected randomly from a standard normal distribution (a mean of 0.0 and a standard deviation of 1.0). Lower asymptote parameters (c's) were selected randomly from a uniform distribution ranging from 0.0 to 0.25. These distributions were selected assuming that the distribution of test takers for the test would be approximately standard normal, such as that on the LSAT. Possible relationships among the IRT parameters were ignored for the purposes of these simulations. This item pool was deemed appropriate to accommodate a 25-item adaptive test, and the necessary Sympson-Hetter simulations (using a standard normal ability distribution of 1,000 test takers) indicated that such a test could be accommodated while satisfying a maximum desired exposure rate of 0.20.

Simulation of Item Theft

Two types of item thieves were considered. Regular item thieves were used to simulate the possibility of regular test takers stealing items and delivering them to future test takers. Professional thieves were used to simulate the possibility of high-ability individuals being hired specifically to steal difficult items.

Regular thieves were simulated by randomly assigning ability values (θ's) from a standard normal distribution. This distribution was selected because regular thieves were assumed to be similar in ability to "regular" test takers, who were assumed to follow a standard normal distribution. (As noted above, a standard normal distribution was used for the distribution of ability among the simulated test takers used for the Sympson-Hetter simulations that determined the exposure parameters that were used for the remaining simulations.)

Professional thieves were assumed to be of substantially greater ability in the domain area. This was simulated by randomly assigning ability values (θ's rom a normal distribution with a mean of 2.5 and a standard deviation of 0.2. Impact of the two types of thieves was evaluated in separate sets of simulations.

Thieves stole items during a standard adaptive test administration. That is, the test was administered to each thief, and responses were based on the 3PL IRT model. Any item seen by a thief was considered stolen and delivered to future test takers (the "coached" test takers). Simulations were conducted for various numbers of thieves separately for each thief type.

Simulation of Test Administration to Coached Test Takers

Once stolen items were identified through simulation of item theft, the behavior of coached test takers (test takers who received the set of stolen items) was simulated. This was done by simulating administration of the adaptive test to a set of test takers with all responses to stolen items designated as correct and using the 3PL IRT model for responses to items that had not been stolen. To ensure accurate estimation of the impact of item theft for test takers of various ability levels, 100 coached test takers were simulated at each of the 13 θ values ranging from −3.0 to 3.0 in increments of 0.5, rather than using test takers from a standard normal ability distribution. To examine the impact on the entire test taker pool, impact at each discrete θ value could be weighted by the proportion of test takers who would fall in each discrete interval, assuming a particular distribution of test taker ability (e.g., a standard normal distribution or a distribution based on empirical findings).

Replication

Simulations were conducted with 2, 5, and 10 regular thieves and 2 and 5 professional thieves. A condition with no thieves (hence no stolen items) was also simulated as a standard for comparison. Each condition was simulated 10 times with a new randomly selected group of thieves for each replication, although the item pool remained constant for all simulations. Reported results are based on averages over the 10 replications for each condition, and the spread is also indicated for some results.

Results

The number and difficulty of the stolen items will be presented first, separately for regular and professional thieves. The impact of item theft on the scores of coached test takers will be shown in several ways, again separately for the conditions using regular and professional thieves.

Number and Difficulty of Stolen Items

In stage one of the simulation study, thieves stole items. This section reports the number and difficulty of the items that were stolen during stage one.

Two regular thieves stole a mean of 46.2 items across the 10 replications (range: 34 to 50 items). Because each thief received a 25-item CAT, the maximum number of stolen items for 2 thieves is 50 items (no overlap between items in the CATs). Fewer than the maximum number of stolen items indicates item overlap among thieves. Five regular thieves stole a mean of 85.8 items across the 10 replications (range: 76 to 99 items;

potential maximum = 125). Ten regular thieves stole a mean of 117.8 items across the 10 replications (range: 90 to135 items; potential maximum = 250 items).

Two professional thieves stole a mean of 31.0 items across the 10 replications (range: 27 to 36 items; potential maximum = 50 items). Five professional thieves stole a mean of 37.1 items across the 10 replications (range: 35 to 41 items; the maximum of 125 items was not reached because of item overlap in the thieves' CATs). Five professional thieves only stole 6.1 more items on average than two professional thieves because of substantial overlap in their CATs. Increasing the number of professional thieves results in only small gains in the number of items that are stolen.

Overall, having more thieves results in more theft, and regular thieves steal more items than professional thieves do. High-ability test takers (and hence professional thieves) receive similar sequences of items when using the unconditional Sympson-Hetter exposure control (Parshall, Davey, and Nering, 1998; Stocking and Lewis,1995) such as was used in the present simulations. Thus, professional thieves had greater item overlap than regular thieves due to increased similarity in ability. Professional thieves also receive a less diverse group of items with this test design as compared to regular thieves. This is shown in Figure 1 which provides the distributions of item difficulty for the items stolen by 5 regular thieves or by 5 professional thieves. Distributions are provided for each of the 10 replications.

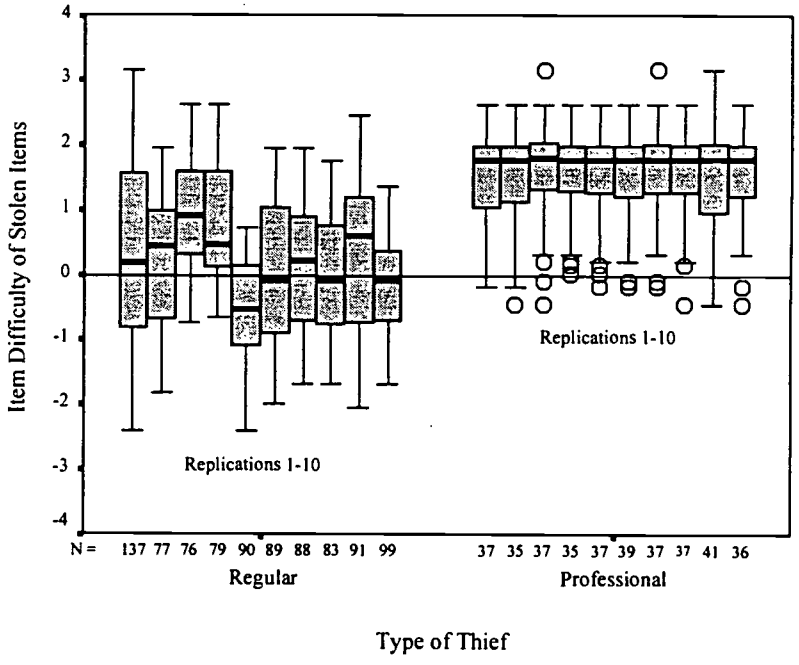

FIGURE 1. *Item difficulty of items stolen by 5 regular thieves and 5 professional thieves in each of 10 replications. N indicates the number of items stolen in each replication. All simulations used the same pool of 500 items.*

Impact of Item Theft on the Scores of Coached Test Takers

In stage two of the simulation study, coached test takers received items in a CAT that they may have seen beforehand (i.e., the stolen items). All test takers in stage two of the simulation were coached. The stolen items were always answered correctly by the test taker. This section reports on the impact of these stolen items on the scores of the coached test takers. Individual ability estimates are first shown at each true ability level for one replication of each condition. Next, the mean ability estimate across replications is shown at each true ability level for each condition. Finally, bias and root mean squared error (RMSE), which are defined below, are shown at each true ability level for each condition.

Individual Ability Estimates at Each True Ability Level

Figure 2 shows $\hat{\theta}$ (estimated ability) by θ (true ability) for coached test takers (stage two of the simulation) when there were no thieves in stage one (Panel A); 2, 5, or 10 regular thieves (Panels B, C, and D, respectively); or 2 or 5 professional thieves (Panels E and F, respectively) for one replication. As shown in Panel A, $\hat{\theta}$ is positively related to θ—low estimates correspond to low true ability and high estimates correspond to high true ability levels—when there are no thieves (i.e., when test takers did not have prior knowledge of any of the items).

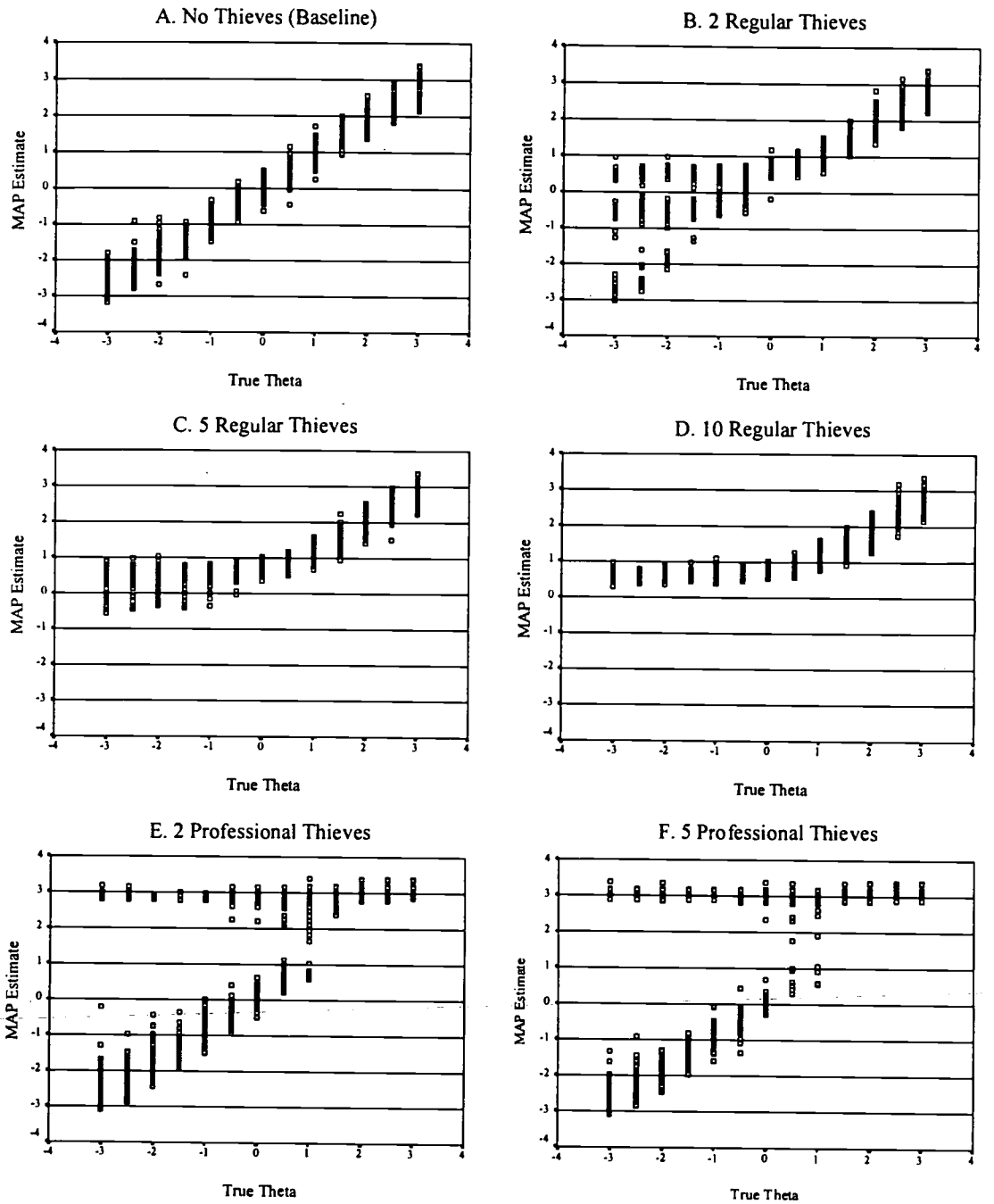

FIGURE 2. *Ability estimate (MAP estimate, or $\hat{\theta}$) by true ability (θ) for one replication of coached test takers in stage two of the simulation when stolen items were provided by no thieves (Panel A); 2, 5, or 10 regular thieves (Panels B-D); and 2 or 5 professional thieves (Panels E-F).*

When the test takers were coached on the items stolen by the thieves, the amount that their ability estimate was impacted was affected by the number of items stolen, the ability level of the test taker, and the ability level of the thieves (and hence the average item difficulty of the stolen items). Panel B of Figure 2 shows that when there were 2 regular thieves, some of the coached test takers with abilities less than 1 were impacted substantially. For example, some coached test takers with $\theta = -3$ received an ability estimate of +1 (4 standard deviations higher than their true ability). Scores for coached test takers with true abilities greater than 1 were not noticeably impacted. As the number of regular thieves increased to 5 and 10 (Panels C and D), all of the coached test takers with abilities less than 1 were noticeably impacted.

When professional thieves were providing the stolen items, the impact on coached test takers' scores followed a different pattern. Because professional thieves had very high ability levels, the items they stole tended to be the most difficult items in the pool. At true ability levels greater than or equal to 1.5, every coached test taker received a substantial number of stolen items and received an estimated score of nearly +3, as shown in Panels E and F. At all ability levels less than or equal to 1, some of the coached test takers had the good fortune of receiving a stolen item as their first or second item. These lucky coached test takers then ended up receiving many stolen items and consequently an estimated ability of about +3. Lower ability coached test takers who did not start with a stolen item ended up seeing few stolen items, and their scores were largely unaffected.

Mean Ability Estimate Across Replications at Each True Ability Level

Figure 3 and Table 1 show the mean estimated ability for coached test takers across the 10 replications for each condition (represented by different lines in Figure 3). As shown in Figure 3, when either 2 or 5 professional thieves provided stolen items, the mean estimated ability was near 3 for coached test takers with true abilities greater than or equal to 1. Test takers with true ability less than one were not helped as much by professional thieves, but their estimated ability still tended to overestimate their true ability. As shown in Figure 2, the mean values in Figure 3 are somewhat misleading for test takers with lower true ability values when professional thieves provide the stolen items because the distribution of estimated abilities for the lower-ability coached test takers is bimodal. Some coached test takers with low true ability values received very high estimated abilities (near 3), whereas the rest of the coached test takers with low true ability values received estimated abilities near their true ability (i.e., they were unaffected by the stolen items).

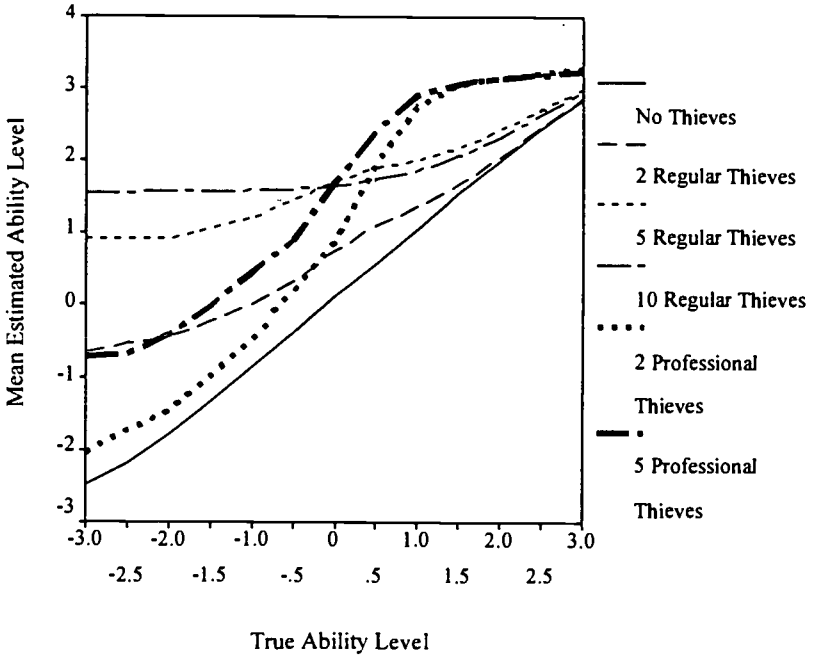

FIGURE 3. *Mean estimated ability level by true ability for coached test takers in stage two of the simulation when stolen items were provided by no thieves (solid line); 2, 5, or 10 regular thieves (thin dashed lines); and 2 or 5 professional thieves (thick dashed lines).*

TABLE 1

Mean estimated ability ($\bar{\hat{\theta}}$) across 10 replications and mean bias (Equation 1) for coached test takers when stolen items were provided by 2, 5, or 10 regular thieves or 2 or 5 professional thieves. The baseline condition in which no stolen items were provided (i.e., no thieves) is also shown. Results are presented separately for each true ability (θ) level.

	No Thieves		2 Regular Thieves		5 Regular Thieves		10 Regular Thieves		2 Professional Thieves		5 Professional Thieves	
True θ	$\bar{\hat{\theta}}$	Mean Bias	$\bar{\hat{\theta}}$	Mean Bias	$\bar{\hat{\theta}}$	Mean Bias	$\bar{\hat{\theta}}$	Mean Bias	$\bar{\hat{\theta}}$	Mean Bias	$\bar{\hat{\theta}}$	Mean Bias
-3	-2.58	0.42	-0.73	2.27	0.81	3.81	1.45	4.45	-2.03	0.97	-0.71	2.29
-2.5	-2.28	0.22	-0.62	1.88	0.81	3.31	1.45	3.95	-1.72	0.78	-0.67	1.83
-2	-1.86	0.14	-0.54	1.46	0.81	2.81	1.46	3.46	-1.45	0.55	-0.41	1.59
-1.5	-1.41	0.09	-0.32	1.18	0.95	2.45	1.46	2.96	-0.96	0.54	-0.02	1.48
-1	-0.95	0.05	-0.08	0.92	1.09	2.09	1.48	2.48	-0.47	0.53	0.45	1.45
-.5	-0.47	0.03	0.23	0.73	1.35	1.85	1.50	2.00	0.09	0.59	0.89	1.39
0	0.00	0.00	0.65	0.65	1.61	1.61	1.53	1.53	0.77	0.77	1.69	1.69
.5	0.47	-0.03	0.99	0.49	1.79	1.29	1.63	1.13	1.85	1.35	2.41	1.91
1	0.95	-0.05	1.26	0.26	1.91	0.91	1.76	0.76	2.66	1.66	2.91	1.91
1.5	1.45	-0.05	1.56	0.06	2.08	0.58	1.96	0.46	2.96	1.46	3.07	1.57
2	1.88	-0.12	1.95	-0.05	2.33	0.33	2.22	0.22	3.04	1.04	3.14	1.14
2.5	2.35	-0.15	2.36	-0.14	2.61	0.11	2.54	0.04	3.11	0.61	3.18	0.68
3	2.75	-0.25	2.76	-0.24	2.89	-0.11	2.85	-0.15	3.18	0.18	3.24	0.24

When regular thieves provided the stolen items, mean estimated abilities for coached test takers were inflated for all but the highest ability levels, especially for 5 and 10 regular thieves, as shown in Figure 3 and Table 1. As noted in Figure 2, with 5 and 10 regular thieves, all coached test takers at a given ability level received similar estimates of ability. Thus, the means in Figure 3 for these test takers are quite representative and are not affected by bimodality.

Bias and Root Mean Squared Error (RMSE) at Each True Ability Level

Bias and RMSE each provides a summary of the impact of stolen items on coached test takers' scores. RMSE provides the average deviation between the estimated and true abilities, regardless of direction, whereas bias provides the average signed deviation. Bias and RMSE were calculated at each θ level.

Bias, the average (signed) deviation between estimated ability ($\hat{\theta}$) and true ability (θ), is given by

$$bias = \frac{\sum_{i=1}^{n}\left(\hat{\theta}-\theta\right)}{n}. \quad (1)$$

where n is the number of test takers at each ability level (100 in the present simulations). A bias value of +2 indicates that ability estimates are 2 standard deviations too high on average (because the θ scale is in standard units).

RMSE, the square root of the average squared deviation between estimated and true ability (the average deviation between $\hat{\theta}$ and θ), is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{\theta}-\theta\right)^2}{n}}. \quad (2)$$

A RMSE value of +2 indicates that ability estimates are 2 standard deviations from the true value on average (in either direction). RMSE emphasizes large deviations more than bias does because the deviations are squared before taking the average.

Bias is shown as a function of true ability (θ) in Figure 4. The condition where there were no thieves (Panel A of Figure 4) provides a baseline against which to compare the effect of having thieves provide stolen item to test takers. As shown in Panel A, there is a slight positive bias for low scores (their ability estimates were slightly higher on average than their true ability), and a slight negative bias for high scores (their ability estimates were slightly lower on average than their true ability). Specifically, as shown in Table 1, for test takers with a true ability of −3, the mean estimated ability was −2.58, and for test takers with a true ability of +3, the mean estimated ability was 2.75. This is a typical result, given the method of estimating ability (Bayes modal estimates generally pull in the tails of the estimated ability distribution slightly).

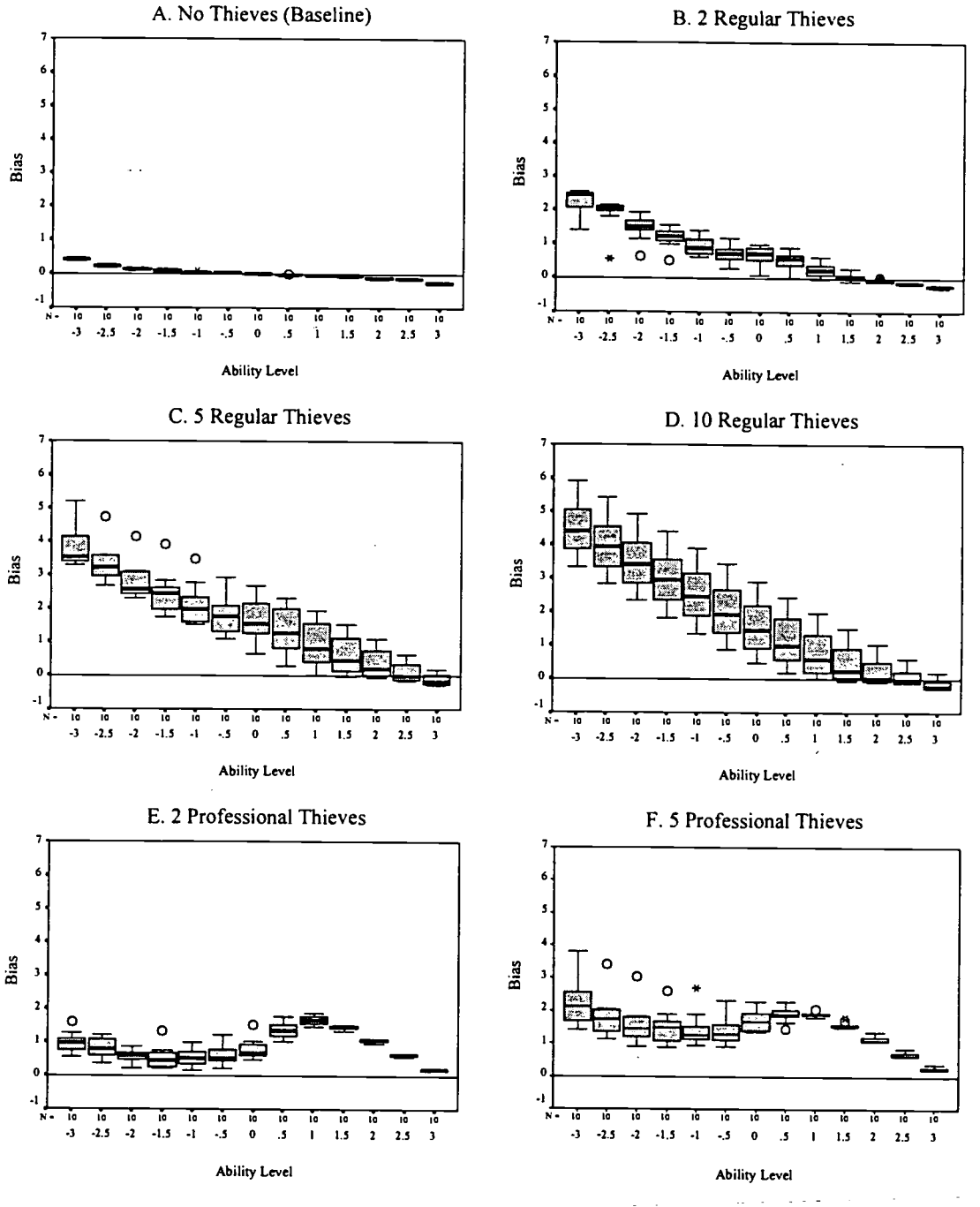

FIGURE 4. *Bias in ability estimates for coached test takers in stage two of the simulation when stolen items were provided by no thieves (Panel A); 2, 5, or 10 regular thieves (Panels B-D); and 2 or 5 professional thieves (Panels E-F).*

Panels B–F of Figure 4 indicate the amount of bias when thieves provided stolen items to test takers. When there were 2 regular thieves (Panel B), bias for all but the highest ability levels increased substantially, compared to the baseline (no thieves). For example, for test takers with a true ability of -3, the mean estimated ability was −0.73, producing an average bias value of 2.27 across the 10 replications, as shown in Table 1. Test takers with a true ability of +3 were unaffected; their mean estimated ability was 2.75, as shown in Table 1, which is the same as in the baseline condition.

As the number of regular thieves increased, bias also increased for all but the highest ability levels. When there were 5 regular thieves (Panel C), test takers with a true ability of −3 had a mean estimated ability of 0.81, producing an average bias value of 3.81 across the 10 replications, as shown in Table 1. When there were 10 regular thieves (Panel D), test takers with a true ability of −3 had a mean estimated ability of 1.45, producing an average bias value of 4.45 across the 10 replications, as shown in Table 1.

The pattern of bias across ability levels when professional thieves supplied the stolen items was different from when regular thieves provided the stolen items, as shown in Panels E and F of Figure 4. Overall bias is lower when professional thieves supplied the stolen items because fewer test takers were affected (e.g., see Figure 2). The amount of bias was very consistent across replications for ability levels greater than or equal to 1 because professional thieves consistently stole the hardest items in the item pool and coached test takers with abilities greater than or equal to 1 consistently received many of these stolen items. Two professional thieves produced average biases of .18, .61, and 1.04 for coached test takers with true abilities of 3.0, 2.5, and 2.0, respectively (see Table 1). Impact for lower ability test takers was less consistent due to the bimodality demonstrated in Figure 2. However, positive bias was demonstrated for test takers at all ability levels.

RMSE (Equation 2) results parallel the bias results very closely, as shown in Figure 5. RMSE, which is an average unsigned difference between estimated and true ability, emphasizes large differences more than bias does, so the RMSE values when stolen items were provided by professional thieves (Panels E and F of Figure 5) are noticeably larger than the corresponding bias values (Panels E and F of Figure 4).

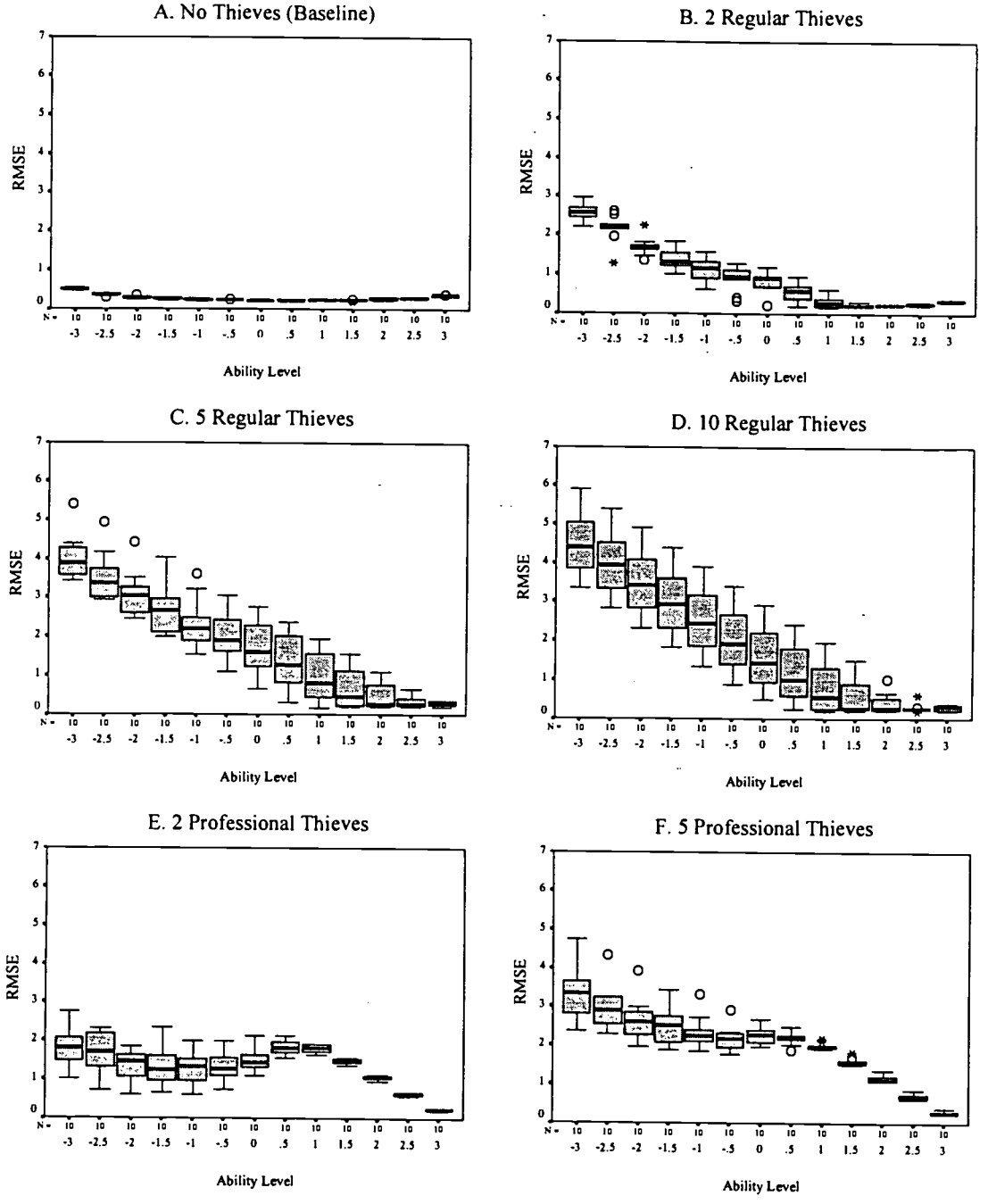

FIGURE 5. *RMSE of ability estimates for coached test takers in stage two of the simulation when stolen items were provided by no thieves (Panel A); 2, 5, or 10 regular thieves (Panels B-D); and 2 or 5 professional thieves (Panels E-F).*

Discussion

In large-scale paper-and-pencil testing, tests are given several times a year. One way that test security is maintained is by carefully controlling when each item appears. Typically, a completely new test form is given at each administration, after which the items are disclosed to the public and are never used again. However, items are usually pretested before being given operationally. Pretest items are sometimes stolen and may be made available to some future test takers. However, future test takers do not know when items will appear again, so

the potential benefit of memorizing stolen pretest items is minimized. In this way, individual item theft is a relatively minor concern in paper-and-pencil testing.

One of the often-mentioned benefits of computer-administered testing is the availability of continuous or on-demand testing. In continuous testing, it would be infeasible to have a completely new item pool for each test taker. If the same item pool is used over time, the possibility that items will become known to the future test takers becomes a problem. The present study explored the hypothetical impact of organized item theft on scores of simulated test takers who were coached on a set of stolen items.

Two types of item thieves were simulated. Regular thieves were of average ability on average, whereas professional thieves were of very high ability on average. Regular thieves stole items with a relatively wide range of difficulty, whereas professional thieves stole primarily difficult items.

Results show that when regular thieves provided the stolen items, all but the highest-ability test takers received inflated ability estimates. When professional thieves provided stolen items, some low-ability test takers were helped tremendously and received ability estimates at the top of the ability range. Additionally, when professional thieves provided the stolen items, all test takers who had relatively high-true abilities received a substantial number of stolen items and also received ability estimates at the top of the ability range.

Although the results may be cause for caution, concerns raised by the current work must be considered in the proper context. For instance, item theft is a concern introduced by continuous testing, not by computer administration. A computer-administered test need not allow continuous testing if all test takers can be accommodated with a set number of prescheduled test administrations. With a set number of administrations, a different item pool could potentially be available for each administration; thus the situation would be similar to current paper-and-pencil administrations.

Another consideration is the appropriateness of assuming perfect memorization. It seems reasonable to assume that thieves can steal items with perfect accuracy because they could use hidden cameras or other technology (Colton, 1997). The questionable assumption is whether test takers could memorize with perfect accuracy the correct answer to numerous stolen items. We felt that this assumption was not unreasonable because most test takers are highly motivated and would probably benefit greatly by selectively memorizing items near (and somewhat above) their ability level, rather than all stolen items.

A third consideration is that the present study investigated one simple test design: a fully adaptive test with constraints only on item exposure. Content constraints were not included, and such constraints could decrease the overlap among the items seen by test takers, so they may be less likely to receive stolen items on their CAT. This would also decrease the overlap among the items seen by thieves, and consequently they would steal more items. The effect of content constraints should be addressed in future research.

Only one item pool was used in the present simulations. One way to counteract the threat of item theft may be to have many separate item pools that are rotated in and out of operational use. Patsula and Steffen (1997) suggested one version of this approach. They suggest a reservoir of all available items from which different item pools are drawn. There is overlap among items in the item pools, but one goal is to keep exposure of individual items to a minimum.

Designs other than a content-balanced CAT with item exposure control may provide more opportunities to counteract the threat of item theft. One potential alternative is the multistage-testlet design currently being considered for possible future computerized LSAT administration (e.g., Reese, Schnipke, & Luebke, 1997; Schnipke & Reese, 1997). In such a design, a prespecified collection of testlets is available in a branching structure. Test takers take various paths through the structure depending on how they perform on each testlet. Hundreds of partially overlapping structures are envisioned, and each structure would be rotated in and out of operational use on some random schedule. On a given day 1-3 structures could be in use. Future research will investigate whether such a scheme would reduce the potential problem of item theft.

The general issue of how to deal with the potential problem of item exposure is currently being investigated by numerous researchers (e.g., Chang, 1998; Kalohn & Spray, 1998; Luecht, 1998; Lunz & Stahl, 1998; McLeod & Lewis, 1998; O'Neill, Lunz, & Thiede, 1998; Tang, Jiang, & Chang, 1998; Thomasson, 1998; Way, 1998). Hopefully, reasonable solutions will be worked out before a serious problem arises.

References

Chang, S-W. (1998, April). *A comparative study of the item exposure control methods in computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Colton, G. (1997, March). *High-tech approaches to breaching examination security.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

Honan, W. H. (1995, January 4). Computer admissions test to be given less often. *The New York Times*, p. A16.

Kalohn, J. C., & Spray, J. A. (1998, April). *Effect of the choice of computerized classification test algorithm on item exposure rates.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Luecht, R. M. (1998, April). *A framework for exploring and controlling risks associated with test item exposure over time.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Lunz, M. E., & Stahl, J. (1998, April). *Patterns of item exposure using a randomized CAT algorithm.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

McLeod, L. D., Lewis, C. (1998, April). *A Bayesian approach to detection of item preknowledge in a CAT.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

O'Neill, T., Lunz, M. E., & Thiede, K. (1998, April). *The impact of item exposure on repeat candidate performance.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Parshall, C. G., Davey, T., & Nering, M. L. (1998, April). *Test dependent exposure control for adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Patsula, L. N., & Steffen, M. (1997, March). *Maintaining item and test security in a CAT environment: A simulation study.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Potenza, M. T., & Stocking, M. L. (1997). Flawed items in computerized adaptive testing. *Journal of Educational Measurement, 34,* 79-96.

Reese, L. M., Schnipke, D. L., Luebke, S. (1997, March). *Incorporating content constraints into a multi-stage adaptive testlet design.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187-208). New York: Springer.

Schnipke, D. L., & Reese, L. M. (1997, March). *A comparison of testlet-based test designs for computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Schnipke, D. L., & Scrams, D. J. (1999). *Representing response-time information in item banks* (Computerized Testing Report 97-09). Newtown, PA: Law School Admission Council.

Stocking, M. L., & Lewis, C. (1995). *Controlling item exposure conditional on ability in computerized adaptive testing.* (Research Report No. RR-95-24). Princeton, NJ: Educational Testing Service.

Stocking, M. L., Ward, W. C., & Potenza, M. T. (1997). *Simulating the use of disclosed items in computerized adaptive testing* (Research Report No. RR-97-10). Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item-exposure rates in computerized adaptive testing.* Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Tang, K. L., Jiang, H., & Chang, H-H. (1998, April). *A comparison of two methods of controlling item exposure in computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

Thomasson, G. L. (1998, April). *CAT item exposure control: New evaluation tools, alternate methods, and integration into a total CAT program.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). *Using response-time constraints in item selection to control for differential speededness in computerized adaptive testing* (Computerized Testing Report 98-03). Newtown, PA: Law School Admission Council.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169-185). New York: Springer.

Wainer, H., Dorans, N. J., Flaugher, R. Green, B. F., Mislevy, R. J., Steinberg, L. & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum.

Way, W. D. (1998, April). *Strategies for managing item pools to maximize item security.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").